LA-UR-19-22515
Approved for public release; distribution is unlimited.

Title:	Short papers on current state of sequencing, metagenomics, and RNAseq for diagnostics
Author(s):	Davenport, Karen Walston
Intended for:	Presentation to sponsors and potential sponsors
Issued:	2019-03-19

Position paper on long and short read sequencing technologies

Long read vs. short read sequencing technologies – considerations when selecting a technology
Chain lab – Los Alamos National Laboratory

Summary Statement: *There are many factors that must be evaluated before ascertaining what type of technology is most cost effective, and/or most technically effective. In general, long read technologies are primarily used for in field/in situ sequencing and for closing isolate genomes, while short read sequencing lends itself to rapid high quality determination of strain identity and to applications that require high depth, such as metagenomes and counting applications, like in RNAseq*.

There are many considerations when establishing what sequencing technology may be used for specific use cases, as there are pros and cons to each technology. Cost estimates are typically shown as a function of a 'run' on the machine, number of nucleotides sequenced within a run, or number of reads obtained from a run. In the latest reports from the sequencing companies, the Illumina MiSeq ($99,000) can produce up to 15 Gigabases of data with 25 million sequencing reads in a single run [1], while the PacBio Sequel ($350,000) can produce up to 10 Gigabases of data with 400,000 sequencing reads [2]; the Nanopore MinION ($2000) can produce up to 30 Gigabases of data (but the number of sequencing reads is not provided) [3]. The estimated costs for sequencing rarely include the cost of the sequencer, maintenance costs, or the time and effort and reagent costs of sample preparation (all of which differ drastically among platforms), the ability to barcode or multiplex samples (loading multiple samples per run), the flexibility for samples (isolates vs. microbiomes, clean *e.g. cultures* vs. dirty *e.g. air filters*, low amount and low yield nucleic acids, RNA vs. DNA capabilities/kits), etc. Therefore, the answer for what technology may be most cost effective depends highly on the specific sample(s) and circumstances.

Run costs and multiplexing capability are important in terms of processing single samples vs. many. The amount of data in terms of number of reads equates to data points, and is important when the use case has counting considerations (for example, more data points are better when looking at counting the number of transcripts in RNAseq, or when enumerating organisms from within microbiome samples). The amount of data in terms of total nucleotides from a run is important for estimating both multiplexing amount or number of runs needed given the goal and complexity and size of genome(s) within the sample. Cost comparisons are thus a very complex issue and are determined by the cross-section of platform, goal of project (depth of genome coverage required), genome or metagenome anticipated size and/or relative abundance, etc. Below, computational algorithm/efficiency is discussed; however, the computational time and resources are an additional important consideration for total cost, yet is rarely factored into the equation.

Outside of cost and the above considerations, the short (Illumina) vs. long (PacBio and Nanopore MinION) technologies have enormous impact on the optimal computational algorithms that should be used and their associated limitations/efficiencies. In general, short reads are limited to under 500 bases (often 100-250 bases for Illumina) and are of high quality (error rates of <<1%) [4], and with Illumina data, all reads are precisely the same length. In contrast, long read technologies have very high error rates (10-15%) [5,6] and read lengths are non-uniform. While long read technologies now provide some data in the >>100 kilobases (kb) range, the average read length is far smaller (10-50kb); read lengths are dependent on the integrity of the sample and the prep.

As a general rule of thumb, in a single run, short read technologies provide much more data in terms of total nucleotides, and orders of magnitude more reads and thus are the only choice for counting

applications (gene expression profiling and enumerating relative abundance of community members within a microbiome). Their high quality lends them well for alignment and the algorithms for this are very efficient in terms of speed and computational memory requirements, which is ideal in read-mapping (alignment) applications.

However, when genome assembly is desired, due to their short length and the nature of biology (conservation of sequences, or duplication and divergence; both within and among genomes), the information content within the short reads (and even in paired reads) is insufficient to resolve repetitive sequences whose length are longer than the reads or the sequenced insert [7,8]. This is the primary reason why genome assembly with short reads, despite fold coverage, does not result in finished genomes. Assembly of unique regions within genomes will still occur most accurately with short read data however. Longer reads can be utilized to help span the repetitive sequence elements and thus long read technologies lend themselves well to assembly applications. Where long reads fail in assembly, it is due to the quality of their data. This is particularly true when using only long read data for assembly. Some recent technologies (e.g. 10X genomics)[9] modify the library creation step for Illumina data processing by isolating long (10's of kb) DNA fragments, creating barcoded libraries from these isolated long fragments, and allowing post-sequencing local assembly of the reads to produce long assembled 'reads' representing these fragments. While these are generally expensive options, they do provide a high quality alternative to long-read technologies, utilizing short Illumina reads. Additional alternative methods, such as using Hi-C crosslinking to co-locate proximal nucleic acids coupled with advanced post-assembly bioinformatics, can also help with assembly efforts, but again are a more expensive option with several pre-processing and post-processing steps [10, 11].

For isolate genomes, where all data is expected to represent a single clonal lineage with the same genome, clever algorithms attempt to 'clean' the errors from the 10-15% error rate by comparing all the data to itself and looking for alignments that suggest they are from the same genomic location. With sufficient coverage of any region in the genome, the algorithm tries to establish the 'correct' sequence at all positions (e.g. with 50-fold coverage, and a 10% error rate, one would expect approximately 5 incorrect sequences at each position in the genome, but 45 correct sequences) [12]. In this manner, long read-only assemblies can achieve a very accurate and complete genome, though the accuracy does not match the accuracy of high quality short read assemblies. The accuracy in these long-read assemblies is also lower in regions where: 1) there is less coverage; 2) there is a local sequencing bias in terms of error; 3) in repetitive regions, since the algorithm 'over-cleans' and normalizes all repetitive sequences to the same sequence [13].

In metagenomes (sequencing complex microbiomes), the algorithms to clean data do not achieve similar performance, since most organisms are not represented by high fold coverage, the organisms are not present in the same abundance, and conserved (repetitive) elements between different genomes can result in genome assembly chimeras.

There are additional, computational considerations when dealing with long read technologies with high error rates. The high error rates (and types of errors – i.e., insertions/deletions) in long reads creates algorithmic challenges in terms of bioinformatic efficiency. These challenges result in increased bioinformatic processing time for read alignment and dramatically increased time and memory requirements for assembly.

A final set of considerations for platform selection are the footprint and lab requirements for the instrument and associated laboratory equipment for sample and library preparation. The Oxford

Nanopore MinION's smartphone-like size and USB-connectivity for power and data capture have allowed the technology to travel to the sample, rather than samples being sent to a centralized laboratory. This is the only technology sufficiently small and robust to be field-deployable.

References:

1. Illumina. (2019). https://www.illumina.com.
2. PacBio. (2019). https://www.pacb.com.
3. Nanopore MinION. (2019). https://nanoporetech.com.
4. Pfeiffer F, Grober C, Blank M, Handler K, Beyer M, Schultze J, and Mayer G. (2018). Systematic evaluation of error rates and causes in short wamples in next-generation sequencing. *Scientific Reports:* 8(10950). doi: 10.1038/s41598-018-29325-6.
5. Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. *Nucleic Acids Research* 46(5), 2159-2168. doi: 10.1093/nar/gky066.
6. Tyler AD, Mataseje L, Urfano C, Schmidt L, Antonation KS, Mulvey MR, and Corbett CR. 2018. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports* 8(10931). doi: 10.1038/s41598-018-29334-5.
7. Scholz MB, Lo CC and Chain, PS. (2012). Next generation sequencing and bioinformatics bottlenecks: the current state of metagenomics data analysis. *Curr Opin Biotechnol*. Feb;23(1):9-15. doi: 10.1016/j.copbio.2011.11.013
8. Hu B, Xie G, Lo CC, Starkenburg SR, and Chain PS. (2011). Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. *Brief Funct Genomics*. 2011 Nov;10(6):322-33. doi: 10.1093/bfgp/elr042.
9. Detailed insight into complex genomes: Our technology uses an informatics method to assemble DNA fragments in to long synthetic reads. (2019). Retrieved from https://www.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html
10. Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo CC, Mueller KA, Sullivan ST, Chain PSG, and Liachko I. (2017). Hi-C deconvolution of a human gut microbiome yields hiqh-quality draft genomes and reveals plasmid genome interactions. *bioRxiv* 198713; doi:10.1101/198713.
11. Burton JN, Liachko I, Dunham MJ, and Shendure J. (2014). *G3: Genes, Genomes, Genetics.* July; 4(7):339-1346. doi: 10.1534/g3.114.011825.
12. Chin CS, Alexander DH, Marks P, Klammer A, Drake J, Heiner C, Clum A, Copeland AC, Huddleston JL, Eichler EE, Turner SW, and Korlach J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature methods.* 10(6). doi: 10.1038/nmeth.2474.
13. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, and Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36,** 338-345. doi: 10.1038/nbt.4060.

	Short reads (Illumina)	Long reads (PacBio)	Long reads (Nanopore MinION)
Maturity of technology	• Stable technology	• Stable Technology	• Nascent, and rapidly advancing technology
Sequencing process	• Sequencing by synthesis using a template of DNA fragments from sample. Fluorophore signals indicate which bases are added in what order.	• Sequencing by synthesis using a template of DNA fragments from sample. Fluorophore signals indicate which bases are added in what order.	• Fragments of DNA are ratcheted through a nanopore 5 bases at a time; electrical signals are recorded and the bases are called using on the pattern of the signal.
Error rate	• Very low error rate (reported error rates are 0.1-2.4%)	• Error rates ~15% (50-100X greater than for short reads).	• Error rates ~15% (50-100X greater that for short reads).
Error types	• Primarily substitutions; some bias	• Primarily random insertions/deletions	• Primarily **non**-random insertions/deletions
Read length (fragment size)	• 300-600 base pairs; all reads equal in length	• >300,000 base pairs; average is 10,000-50,000 bp	• >>2,000,000 base pairs; average is 1000-10,000 bp
DNA / RNA	• RNA must be converted to cDNA for sequencing	• RNA must be converted to cDNA for sequencing	• Can read both DNA and RNA (without converting RNA to cDNA).
Lab protocols	• Stable protocols	• Stable protocols	• Changing as technology matures
Analytical tools	• Thousands of bioinformatic pipelines for analysis	• A handful of proprietary pipelines are the primary methods of analysis	• Limited, but growing as technology matures
Versions of sequencer	• Production Scale: NovaSeq, HiSeq, NextSeq • Benchtop: NextSeq, MiSeq, MiniSeq, iSeq	• RSII, Sequel	• R9.4, R9.4 Spot-on, R9.5, RevD flow cells; Flongle, MinION • Flow cells arrayed for higher throughput: GridION, PromethION
Portability	• NovaSeq and HiSeq are large floor models; all others are benchtop. MiSeq, MiniSeq, and NextSeq are roughly the size of a microwave oven, while the iSeq is fairly small- the size of a toaster.	• Both the RSII and the Sequel are large floor models.	• The sequencer itself is field-deployable (~size of an iPhone). The library prep equipment to process the sample have been significant; however, recent improvements have reduced these requirements.
Base calling software	• Cloud or software on-site	• Software on-site	• Cloud or software on-site
Advantages	• Low error rate provides much higher confidence in isolate strain characterization, SNP/INDEL analysis, and taxonomic classification from metagenomic samples. • Can utilize very low DNA input. • Can utilize degraded samples.	• Longer reads mean more complete assemblies with fewer reads since repeat regions are generally spanned by the reads. • Full length transcripts and splice variants are attainable. • Can detect DNA modifications.	• Longer reads mean more complete assemblies with fewer reads since repeat regions are generally spanned by the reads. • Full length transcripts and splice variants are attainable. • Can detect DNA modifications.
Disadvantages	• Draft assemblies are expected since repeat regions are not spanned by the reads. Some recent technologies can create synthetic long reads with short read sequencing platforms. While this is an expensive option, it does provide some solutions to the issues with short reads (i.e., allelic variations or separating organisms in a metagenomic sample).	• High error rate makes some types of analyses more challenging and less reliable, but this is overcome by high throughput providing high depth of coverage to correct errors. • Requires high concentration of high-quality DNA. • Plasmids shorter than the average read length may be difficult to detect or assemble depending on data QC and assembly tools selected.	• High error rate makes some types of analyses more challenging and less reliable. • Requires high concentration of high-quality DNA. • Inconsistencies in throughput per run. • Plasmids shorter than the average read length may be difficult to detect or assemble depending on data QC and assembly tools selected. • So new and rapidly changing, there has not been enough benchmarking.

Position paper 3: RNAseq for diagnostics

RNAseq and implications for diagnostics

Chain lab – Los Alamos National Laboratory

Summary statement: *RNAseq can provide some important insights into the biology of active infections and provide advances in diagnostics approaches. Depth of sequencing coverage is critical to determine gene expression profiles. Full use of RNAseq data from both host and pathogen(s) for diagnostics and informed treatment options will require further investigation.*

RNA sequencing (RNAseq) targets the pool of transcribed nucleic acids instead of the DNA fraction and is most often used in targeting RNA viruses, as well as to describe gene expression changes/patterns in living organisms. Its use in pathogen detection and for host biomarker expression towards diagnostics has also been explored in limited fashion. For pathogen detection and/or host-response purposes, sample collection (of the correct sample type) timing and preparation are important considerations, together with pathogen load and potential for background noise.

Outside of sequencing RNA viruses, one of the benefits of sequencing RNA instead of DNA is to identify and understand what genes and what organisms are actively expressing RNA (i.e., are alive and active) within the sample at a specific timepoint. For this reason, it has even been used by our group to identify pathogens within clinical samples, given the assumption that infectious pathogens are expressing a number of genes (including virulence factors) during disease progression. We have even recovered entire viral genomes using this approach. While we have shown that pathogen RNAs for DNA and RNA viruses, as well as bacterial pathogens can be detected in a more sensitive fashion using RNAseq compared with the same sample undergoing DNA sequencing, the sequences detected represent only a small, highly expressed, fraction of the pathogen genome (e.g. rRNA), limiting its utility of characterizing the organisms down below the genus or species level.

Removal of non-informative RNAs (particularly rRNAs that can make up to 80% of total RNA [1], and for blood samples, globin mRNA [2]) is thus essential to both increase the sensitivity of this approach and provide a finer-scale resolution in terms of both pathogen identity and host-pathogen response (i.e., gene expression pattern). When non-informative RNAs are depleted, the goal of RNAseq in diagnostics is to:

1. establish and quantify gene expression levels for the pathogen, allowing resolution of within-species discrimination of invasive vs. commensal strains, and providing signatures of antibiotic resistance, etc.
2. establish and quantify gene expression levels of the host in response to the invading pathogen (i.e., RNA biomarkers), in order to discriminate pre-symptomatic gene expression profiles of early infection versus later gene expression responses

While pathogen detection in clinical samples will be limited in the same fashion as in traditional metagenomics sequence to taxonomic identity analysis (*see position paper on the state of*

metagenomic sequencing), because the RNAseq approach targets expressed genes, the distribution of sequencing data around the target genome will be unevenly distributed and will provide an incorrect assessment of pathogen abundance quantification (instead, the signal from RNAseq should simply be interpreted as gene expression quantification). Taxonomy identification will be limited by the number of genes being expressed (and later sequenced) at the time of sampling. Pathogens with reference genomes in the database should be readily identified, while novel pathogens will suffer the same issues as described in the *position paper on the state of metagenomic sequencing*. For the host however, sequences will be mapped to the representative genome, and gene expression quantification can be compared with gene expression profiles of many other pathogens in the same host (there exist a number of studies, particularly in human cell lines, in mice, and other animals, infected with viral and bacterial pathogens) [3]. These will form the basis for understanding if the host response itself is sufficient to be able to 1) determine if the host is infected with a pathogen, 2) what infection control measures may be useful to implement, 3) what treatment may be effective against the pathogen, 4) what the outcome may be with/without treatment, etc. This potential outcome of RNAseq as a diagnostic would greatly impact the current state of prophylactic treatment in public health and warfighter settings.

Full realization of this type of RNAseq for diagnostics remains a long-term vision, as much data is required to be collected before such predictive analyses can be successfully made. Given sufficient data and examples that may account for individual host-specific unique responses, pathogen strain variation, differences in response based on host or environmental factors (time of day, temperature, male/female, what the original gut microbiome composition is/was, what foods were ingested prior to sampling, etc.), a machine-learning framework may be able to help us develop an infection 'classifier' that can help triage individuals both pre- and post-symptomatically, and determine what might be best to do next.

Regarding technologies, while the high error rate in long reads may affect pathogen identification, longer reads are able to be mapped efficiently to the host genome, and may provide information regarding differential gene-splicing events during infection. The primary drawback to long read technologies compared with short read technologies in this case, similar to metagenomics, is that the superior number of reads from short read technologies is of greater importance than read length for quantifying gene expression levels. *(Also see position paper on long and short read sequencing technologies).*

References

1. Condon, C., *Maturation and degradation of RNA in bacteria.* Curr Opin Microbiol, 2007. **10**(3): p. 271-8.
2. Vartanian, K., et al., *Gene expression profiling of whole blood: comparison of target preparation methods for accurate and reproducible microarray analysis.* BMC Genomics, 2009. **10**: p. 2.
3. Westermann, A.J., L. Barquist, and J. Vogel, *Resolving host-pathogen interactions by dual RNA-seq.* PLoS Pathog, 2017. **13**(2): p. e1006033.

RNAseq for Diagnostics

Advantages	Disadvantages
• Detects all **active** pathogens since RNA degrades quickly within host after pathogen death. • Only way to detect pathogenic RNA viruses. • Can provide improved sensitivity for detection using abundant pathogen transcripts, and/or host response genes. • Can provide quantification of functional gene expression levels for the pathogen (antibiotic resistance, virulence, etc.) • Can provide quantification of gene expression levels for the host response (RNA biomarkers), perhaps pre-symptomatically.	• Requires special sample prep to preserve RNA at sample collection since RNA is less stable than DNA. • Need to remove non-informative host RNA to enrich for both pathogen RNA and informative host RNA. • Need further studies of time series for host-pathogen interactions and gene expressions for both host and pathogen in various types of infections. • Need further investigation into a multitude of variables which affect host-pathogen interactions (i.e., host, pathogen strain, and environmental factors) for full utilization of RNAseq.

Position paper 3: RNAseq for diagnostics

RNAseq and implications for diagnostics

Chain lab – Los Alamos National Laboratory

Summary statement: *RNAseq can provide some important insights into the biology of active infections and provide advances in diagnostics approaches. Depth of sequencing coverage is critical to determine gene expression profiles. Full use of RNAseq data from both host and pathogen(s) for diagnostics and informed treatment options will require further investigation.*

RNA sequencing (RNAseq) targets the pool of transcribed nucleic acids instead of the DNA fraction and is most often used in targeting RNA viruses, as well as to describe gene expression changes/patterns in living organisms. Its use in pathogen detection and for host biomarker expression towards diagnostics has also been explored in limited fashion. For pathogen detection and/or host-response purposes, sample collection (of the correct sample type) timing and preparation are important considerations, together with pathogen load and potential for background noise.

Outside of sequencing RNA viruses, one of the benefits of sequencing RNA instead of DNA is to identify and understand what genes and what organisms are actively expressing RNA (i.e., are alive and active) within the sample at a specific timepoint. For this reason, it has even been used by our group to identify pathogens within clinical samples, given the assumption that infectious pathogens are expressing a number of genes (including virulence factors) during disease progression. We have even recovered entire viral genomes using this approach. While we have shown that pathogen RNAs for DNA and RNA viruses, as well as bacterial pathogens can be detected in a more sensitive fashion using RNAseq compared with the same sample undergoing DNA sequencing, the sequences detected represent only a small, highly expressed, fraction of the pathogen genome (e.g. rRNA), limiting its utility of characterizing the organisms down below the genus or species level.

Removal of non-informative RNAs (particularly rRNAs that can make up to 80% of total RNA [1], and for blood samples, globin mRNA [2]) is thus essential to both increase the sensitivity of this approach and provide a finer-scale resolution in terms of both pathogen identity and host-pathogen response (i.e., gene expression pattern). When non-informative RNAs are depleted, the goal of RNAseq in diagnostics is to:

1. establish and quantify gene expression levels for the pathogen, allowing resolution of within-species discrimination of invasive vs. commensal strains, and providing signatures of antibiotic resistance, etc.
2. establish and quantify gene expression levels of the host in response to the invading pathogen (i.e., RNA biomarkers), in order to discriminate pre-symptomatic gene expression profiles of early infection versus later gene expression responses

While pathogen detection in clinical samples will be limited in the same fashion as in traditional metagenomics sequence to taxonomic identity analysis (*see position paper on the state of*

metagenomic sequencing), because the RNAseq approach targets expressed genes, the distribution of sequencing data around the target genome will be unevenly distributed and will provide an incorrect assessment of pathogen abundance quantification (instead, the signal from RNAseq should simply be interpreted as gene expression quantification). Taxonomy identification will be limited by the number of genes being expressed (and later sequenced) at the time of sampling. Pathogens with reference genomes in the database should be readily identified, while novel pathogens will suffer the same issues as described in the *position paper on the state of metagenomic sequencing*. For the host however, sequences will be mapped to the representative genome, and gene expression quantification can be compared with gene expression profiles of many other pathogens in the same host (there exist a number of studies, particularly in human cell lines, in mice, and other animals, infected with viral and bacterial pathogens) [3]. These will form the basis for understanding if the host response itself is sufficient to be able to 1) determine if the host is infected with a pathogen, 2) what infection control measures may be useful to implement, 3) what treatment may be effective against the pathogen, 4) what the outcome may be with/without treatment, etc. This potential outcome of RNAseq as a diagnostic would greatly impact the current state of prophylactic treatment in public health and warfighter settings.

Full realization of this type of RNAseq for diagnostics remains a long-term vision, as much data is required to be collected before such predictive analyses can be successfully made. Given sufficient data and examples that may account for individual host-specific unique responses, pathogen strain variation, differences in response based on host or environmental factors (time of day, temperature, male/female, what the original gut microbiome composition is/was, what foods were ingested prior to sampling, etc.), a machine-learning framework may be able to help us develop an infection 'classifier' that can help triage individuals both pre- and post-symptomatically, and determine what might be best to do next.

Regarding technologies, while the high error rate in long reads may affect pathogen identification, longer reads are able to be mapped efficiently to the host genome, and may provide information regarding differential gene-splicing events during infection. The primary drawback to long read technologies compared with short read technologies in this case, similar to metagenomics, is that the superior number of reads from short read technologies is of greater importance than read length for quantifying gene expression levels. *(Also see position paper on long and short read sequencing technologies).*

References

1. Condon, C., *Maturation and degradation of RNA in bacteria.* Curr Opin Microbiol, 2007. **10**(3): p. 271-8.
2. Vartanian, K., et al., *Gene expression profiling of whole blood: comparison of target preparation methods for accurate and reproducible microarray analysis.* BMC Genomics, 2009. **10**: p. 2.
3. Westermann, A.J., L. Barquist, and J. Vogel, *Resolving host-pathogen interactions by dual RNA-seq.* PLoS Pathog, 2017. **13**(2): p. e1006033.

RNAseq for Diagnostics

Advantages	Disadvantages
• Detects all **active** pathogens since RNA degrades quickly within host after pathogen death. • Only way to detect pathogenic RNA viruses. • Can provide improved sensitivity for detection using abundant pathogen transcripts, and/or host response genes. • Can provide quantification of functional gene expression levels for the pathogen (antibiotic resistance, virulence, etc.) • Can provide quantification of gene expression levels for the host response (RNA biomarkers), perhaps pre-symptomatically.	• Requires special sample prep to preserve RNA at sample collection since RNA is less stable than DNA. • Need to remove non-informative host RNA to enrich for both pathogen RNA and informative host RNA. • Need further studies of time series for host-pathogen interactions and gene expressions for both host and pathogen in various types of infections. • Need further investigation into a multitude of variables which affect host-pathogen interactions (i.e., host, pathogen strain, and environmental factors) for full utilization of RNAseq.